DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013

How Portable Are the Metadata Standards for Scientific Data? A Proposal for a Metadata Infrastructure

Jian Qin
School of Information Studies
Syracuse University
USA
jqin@syr.edu

Kai Li
School of Information Studies
Syracuse University
USA
kli18@syr.edu

Abstract

The one-covers-all approach in current metadata standards for scientific data has serious limitations in keeping up with the ever-growing data. This paper reports the findings from a survey to metadata standards in the scientific data domain and argues for the need for a metadata infrastructure. The survey collected 4400+ unique elements from 16 standards and categorized these elements into 9 categories. Findings from the data included that the highest counts of element occurred in the descriptive category and many of them overlapped with DC elements. This pattern also repeated in the elements co-occurred in different standards. A small number of semantically general elements appeared across the largest numbers of standards while the rest of the element co-occurrences formed a long tail with a wide range of specific semantics. The paper discussed implications of the findings in the context of metadata portability and infrastructure and pointed out that large, complex standards and widely varied naming practices are the major hurdles for building a metadata infrastructure.

Keywords: metadata standards; scientific data; metadata infrastructure.

1. Introduction

Elements in metadata standards for scientific data convey the essential information about the creators, contexts, geospatial and temporal parameters, quality details concerning a dataset or collection. These entities – creator, dataset, geolocation, temporal measurement, instrument, etc. – form an interrelated network of nodes with different functions in supporting the data discovery, selection, locating, and obtaining tasks for data users and the data organization, management, and preservation for data managers. Representing the characteristics of all these entities and their relationships in the form of metadata proves to be a daunting task. One just needs to take a look at the sheer numbers of elements in metadata standards in science domains and the complicated linguistic and syntactic forms used in these standards to understand the magnitude of the challenge.

At present most metadata standards for scientific data have focused on describing datasets, usually the data products at the end of a research lifecycle. As such, they are required to fulfill the goals of preservation, interoperability, reuse, and sharing of data (Jones et al., 2001; Michener, 2006), which necessitates a long list of categories of metadata to support the goals of metadata for scientific data. These requirements make metadata standards for science data inherently complex and need highly trained professionals to created standards-conformed metadata records. Creating a record using the Content Standard for Digital Geospatial Metadata (CSDGM) standard, for example, can take hours to complete due to the large number (300+) of element values that need to be entered manually. In the face of exponential growth of scientific data, the slow pace of metadata creation is making the conventional approach used in developing metadata standards out of date.

Practitioners and researchers have long realized the limitations of science metadata standards. Riall et al. pointed out that "the FGDC-CSGM is so specific that it becomes unwieldy to apply, and thus is undesirable for a catalogue of Web-based materials which are not necessarily raw data and for which much of the information required by the FGDC-CSGM may not be readily available

https://doi.org/10.23106/dcmi.952136097

or desirable for searching and browsing" (2004, section 4). Another important standard, Ecological Metadata Language (EML), is considered as complex and interpretively flexible, which "makes it difficult to understand and enact in its entirety and requires a complete redesign of their data management structures and practices" (Millerand & Baker, 2010, p. 146).

While conventional metadata approaches to representing and managing scientific data are encountering difficulties to keep up with the fast growth of scientific data, new technologies and practices are emerging in cyberinfrastructure-supported, data-driven science and making way for unconventional metadata development and deployment. Technical standards such as Resource Description Framework (RDF), Gleaning Resource Descriptions from Dialects of Languages (GRDDL), RDFa, and Linked Data are providing or will provide the capacity for a much more granular data representation on the Web. Standards important for a metadata infrastructure have also been developed, e.g., Open Researcher and Contributor ID (ORCID), Digital Object Identifier (DOI), ResearcherID, and Uniform Resource Identifier (URI). Clearly we are at a juncture where a tremendous need for metadata infrastructure services meets the enabling technologies. What action should and can we take at this juncture as a community of metadata practices? How much do we know about metadata standards for scientific data? How can we transform the current metadata standards into an infrastructure-driven service? These are the key questions we need to address if we were to significantly improve the metadata services for scientists and scientific data.

In this paper, we report preliminary findings from a survey of metadata standards for scientific data. This study is an attempt to building some groundwork for addressing the three questions mentioned above. The survey focused on the metadata elements: we examined what elements are common across different standards, into what categories these metadata elements can be grouped, and what linguistic, syntactic, and semantic features exist in the elements. Finally, how these features might contribute to the portability of metadata standards.

2. A Metadata Infrastructure for Scientific Data

The word "infrastructure" refers to the underlying foundation or basic framework of a system or organization in a locality or country. Infrastructure is embedded in or inside of other structures, social arrangements, and technologies and does not need to be invented each time or assembled for each task. It reaches or scopes beyond a single event or a local practice and is learned as part of membership while links with conventions of practice. Infrastructure is the embodiment of standards and built on an installed base. It is fixed in modular increments, not all at once or globally (Leigh Star and Ruhleder, 1996). The general features of infrastructure also apply to the scenario of a metadata infrastructure.

Metadata infrastructure by definition implies that metadata elements, vocabularies, entities, and other metadata artifacts are established as the underlying foundation upon which the tools and applications as well as functions of metadata services are built. As early as in 2004, the concept of metadata infrastructure was used by Roy Tennant in the context of bibliographic metadata infrastructure. Tennant maintains that a bibliographic metadata infrastructure should have: 1) versatility, 2) extensibility, 3) openness and transparency, 4) low threshold, high ceiling, 5) cooperative management, 6) modularity, 7) hierarchy, 8) granularity, and 9) graceful in failure (Tennant, 2004). The last decade saw ongoing efforts in building a metadata infrastructure in the library community, ranging from the metadata translation services at OCLC (Godby et al., 2003) to the joint effort between the Library of Congress (LC), National Agricultural Library (NAL), and the National Library of Medicine (NLM) in building a robust metadata infrastructure for the future that are embodied in the RDA Toolkit (U.S. RDA Test Coordinating Committee, 2013).

There are, however, varying uses of the term "metadata infrastructure." In a report for the Common Language Resources and Technology Infrastructure (CLARIN), Wittenburg considers the technical infrastructure as the actual existing implementation of a metadata infrastructure that should include a data model for describing the resources, aspects of metadata encoding and

https://doi.org/10.23106/dcmi.952136097

storage formats, metadata for web services, metadata tools, usage, modification, transformation, interoperability, and metadata crosswalk (Wittenburg, 2009). CLARIN is not the only one holding the technical architecture view on metadata infrastructure. The National Cancer Institute's caGrid system is described as "a service-oriented platform that supports cutting-edge collaborative e-Science by providing the tools for organizations to integrate data silos" (http://www.cagrid.org/display/cagridhome/Home). The caGrid metadata infrastructure includes a collection of components: standardized service metadata models, metadata model services that generate and semantically annotate standard metadata, global model exchange that acts as the authoritative repository for XML schemas used on the grid, index service, Cancer Standard Data Repository (caDSR), Enterprise Vocabulary Services (EVS), and discovery (caGrid, 2011). Another project, the Earth System Curator, developed a metadata formalism for describing the digital resources used in climate simulations, which was called a "metadata infrastructure for climate modeling" (Dunlap et al., 2008).

These developments suggest at least two components – the semantic and technical components – in a metadata infrastructure. The semantic component includes the vocabularies, schemas, and models that consist of standards for metadata description. The technical component takes the semantic component into action – operate and deliver metadata services. There is a third dimension in this metadata infrastructure picture, that is, the policy component that encompasses best practice guidelines, rules, and policies governing the metadata practice. While the metadata infrastructure is not a new concept – it has been around for at least a decade, the current development in technologies and digital data offers a new perspective to re-examine the concept, especially in the context of scientific data. Ten years ago the Semantic Web technologies seemed to be far out of reach. Ten year later many resources for building a metadata infrastructure have become openly accessible – the Library of Congress Subject Headings (LCSH), DBpedia, the largest open linked data repository, and the discipline-oriented metadata infrastructures mentioned earlier in this paper – to name only a few. The changing landscape of technologies and digital data is challenging the conventional approach in designing metadata schemas and applications. Clearly, we need to re-examine those very large, comprehensive metadata standards for scientific data to see how they can be adapted to this changing technology and data landscape in order to build a metadata infrastructure that can more effectively operate and deliver metadata services.

3. Data Collection

The data collection was based on an overarching research question of this study: how portable are the metadata standards for scientific data? The rationale for this research question is that smaller, portable metadata standards are the organic components of a metadata infrastructure and allow for flexible assembling of description models. A survey of metadata standards in the science domain will provide the understanding of their semantic, structural, and contextual attributes, which is the very foundation for addressing the portability question. A similar study by Willis et al. (2012) collected 9 metadata schemes used in active data repositories to study the scope, similarities and/or dissimilarities of metadata schemes. The goal of their study was to derive fundamental requirements of metadata schemes for scientific data. Compared to Willis et al.'s study, a significant difference is that this project focuses on the semantic and linguistic patterns of elements in standards and the purpose is to uncover how portable these elements might be for building a metadata infrastructure.

The portability of metadata standards in the context of this study is defined as the ability of semantic elements in a metadata standard to be reused in different contexts through interoperable applications. Two measures can be applied to assess the portability of metadata elements:

Co-occurrence of semantic elements: Elements that are semantically identical but may vary linguistically are used in multiple standards. The assumption is that the more frequently a

https://doi.org/10.23106/dcmi.952136097

semantic element co-occurs in multiple standards, the more likely it is for the element to be reused in cross-domain representation.

Degree of modularity: Elements representing one concept/entity may be organized as a group or cluster that contains a self-explanatory sub-structure. Such a sub-structure can be an independent part of the whole standard. The more concepts/entities in a standard are organized with independent, self-explanatory sub-structures, the higher the degree of modularity.

The data collected for this project focused on these two portability measures. Table 1 contains a list of metadata standards included in this study. Elements were extracted from all the standards in Table 1 and entered into a Microsoft Access database. For standards with XML schema files, XSLT programs were used to extract elements from these schemas automatically. When a standard contained only natural language specification (without an XML schema), the elements were entered into the database manually. Two standards listed neither schema nor specification, including ClinicalTrials.gov Protocol Data Element Definitions and Genome Metadata. In this case, other documents were sought as reference specifications in order to collect the elements. However, the document of Genome Metadata we used included only the list of elements without any explanation of these elements. The categorization of the elements in Genome Metadata was mainly based on the semantics as reflected in the element names.

TABLE 1. Metadata standards for scientific data included in the study

Title	Abbr.	Website	Date	Version
Access to Biological Collection Data	ABCD	http://www.tdwg.org/	2005	2.06
Astronomy Visualization Metadata Standard	AVM	http://www.virtualastronomy.org/AVM_DRAFTVersion1.1_rlh27.pdf	2008	1.10
ClinicalTrials.gov Protocol Data Element Definitions	Clinical	http://www.nlm.nih.gov/	2012	Draft
Content Standard for Digital Geospatial Metadata	CSDGM	http://www.fgdc.gov/	1998	2.0
Content Standard for Digital Geospatial Metadata, Part 1: Biological Data Profile	CSDGM-BD	http://biology.usgs.gov/cbi/	1999	2.0
Darwin Core	Darwin	http://www.tdwg.org/	2006	2.0
Dublin Core Metadata Element Set	DC	http://dublincore.org/	2008	1.1
Ecological Metadata Language	EML	http://www.nceas.ucsb.edu/	2009	2.1.1
GenBank Flat File Format	GenBank	http://www.ncbi.nlm.nih.gov/		
Genome Metadata	Genome	http://enews.patricbrc.org/faqs/genome-metadata-faqs/		
International Virtual Observatory Alliance	IVOA	http://wiki.ivoa.net/twiki/bin/view/IVOA/IvoaResReg#IVOA_Resource_Registry_Working_G	2007	1.12
ISO/TS 19115:2003 Geographic information — Metadata	ISO 19115	http://www.iso.org/iso/home.html	2003	
Metadata Profile for Shoreline Data	CSDGM-SL	http://www.fgdc.gov/participation/working-groups-subcommittees/mcsdsc	2001	001.2
NetCDF Climate and Forecast Metadata Convention	CF	http://cf-pcmdi.llnl.gov/	2011	1.6
NISO Metadata for Images in XML	NISO-Image	http://www.loc.gov/marc/ndmso.html	2008	2.0
WHO Trial Registration Data Set	TRDS	http://www.who.int/ictrp/network/trds/en/index.html		1.2.1

A total of 5,800 elements were gathered from the 16-metadata standards and grouped into nine categories, in which 4434 elements were unique based on the form. Each category was given a scope note to define the boundaries of that category as well as the guiding principles for element categorization (see Table 2).

https://doi.org/10.23106/dcmi.952136097

TABLE 2. Description of the category scope

Category	Scope of the category
Administrative	• Meta-metadata, i.e., information about metadata record, standard used, responsible party, rights for the metadata record, etc. • Information about data archive/repository.
Context	• Information about study/project design, model, and population under study. • Data collection methods, instruments, and constraints. • Analysis methods used.
Descriptive	• General attributes about what the resource is and when it is published, released, or made available. • Related resources of the resource that is described.
Geospatial	• Geographic names. • Geospatial coordinates. • Aerial maps and/or data.
Generic	• General-purpose elements, including comment, annotation, note, etc. • Wrapper or nesting elements for structuring and syntactic purposes.
Identity	• The name of an entity that is used to identify the entity understood by human users. • A unique ID either in the form of some code or of a string following an identification system.
Semantic	• Subject terms describing the content of data. • Subject or classification categories. • Taxonomic classes.
Temporal	• Measurements of time. • Temporal coverage of the content of data. • Temporal criteria for data segmentation, processing.
Technical	• Parameters, models, measurements used in the dataset. • Software-, system-, and format-related attributes.

In the process of categorization, there were situations where an element may be grouped into two categories. For example, the element "extent" is defined in ISO 19115:2003 as the spatial, horizontal and/or vertical, and the temporal coverage in the resource. In this case, it was assigned to both temporal and geospatial categories. If the same element could be grouped into two categories but one category was the primary and other only the minor aspect of the element, only the primary category was assigned to the element. Similarly, elements playing identification role only locally were not classified to the identify category, but rather, they were grouped into the technical category. An example would be the ID-in-Database element in ABCD.

The categorization of elements was checked between the two coders and uncertainties due to subtle differences in semantic meanings were resolved through discussion and referencing standards documents.

Once the categorization was finalized, frequencies for categories and elements were calculated for the first round of analysis and data cleaning. At this stage, data were sorted by different orders and criteria to aggregate elements that were semantically same but linguistically different. The purpose was to make sure that elements with the same semantic meanings were brought together as one "super-element" regardless of which standard it came from, under which context an element appeared, or whether an element used singular or plural form, lower or capital case, or different word sequences.

4. Findings

4.1 General Description

The overall frequency distribution of element categories in FIG. 1 contained a distinct pattern: three categories – semantic, generic, and temporal – situated at the lower end, four categories ranged in the 400s, while the higher end were occupied by geospatial and context categories. The extremely large number of context elements prompted a further examination of the elements in this category, which revealed that not all metadata standards for scientific data place the focus of

https://doi.org/10.23106/dcmi.952136097

description on *datasets*. The extremely large number of elements in the context category was mainly due to one standard: the NetCDF Climate and Forecast (CF) Metadata Conventions. The NetCDF CF provides a definitive description of what the data in each variable represents as well as the spatial and temporal properties of the data. In other words, this standard represents data at the variable level, rather than at the dataset level as most other standards do.

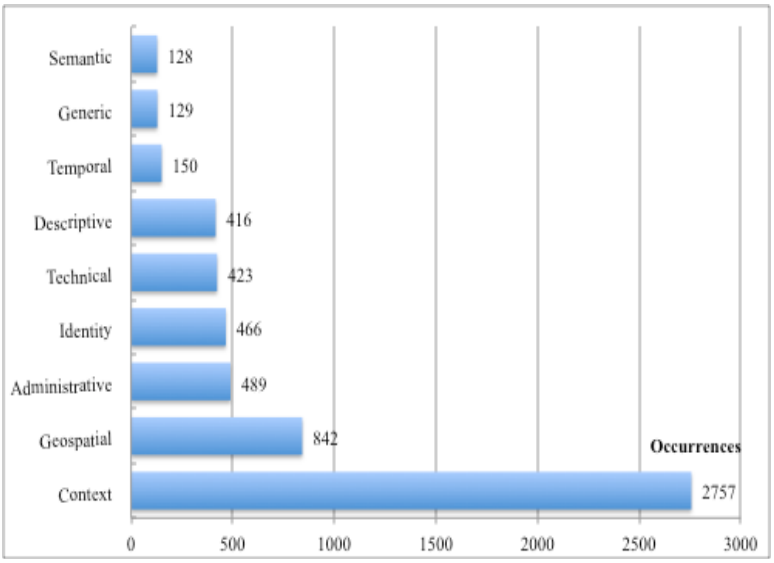

FIG. 1. Frequency of occurrences by category

The top 8 element occurrences all happened in the descriptive, generic, identity, and technical categories (Table 3). The two largest numbers (19 & 17) are greater than the number of standards, which indicate that these elements were used more than once in some of the metadata standard. It is interesting that almost all these categories and elements match the ones commonly used in general metadata standards such as Dublin Core and DataCite.

TABLE 3. Top occurring elements

Category	Element name	Counts
Descriptive	Description	19
Descriptive	Citation	17
Descriptive	Publisher	14
Descriptive	Comment	14
Descriptive	Publication	10
Descriptive	Reference	10
Descriptive	Title	10
Technical	Method	9
Identity	Name	9
Identity	Address	8
Technical	units	8
Descriptive	abstract	7
Generic	attributeList	7
Descriptive	edition	7
Identity	Identifier	7
Descriptive	source	7
Descriptive	Purpose	7

4.2 Elements inside Categories

After obtaining the overall picture of element categories, we turned to analyze the element names in individual categories. The top occurring elements in the identity category included Name (9), Address (7), attributeReference (5), City (5), Identifier (5), Country (4), Email (4), GenusOrMonomial (4), Key (4), Phone (4), and Postal Code (4). The elements in this category contained a wide variety of element names for the same or different semantic meanings. For example, a large number of elements for the roles (e.g., creator, imageProducer, investigators) and contact information of persons and organizations appeared only once. There were also a large number of elements for various entities, including organism, project, taxon, facility, gene, collection, and dataset. When aggregating these elements by their semantic meaning, the data showed 8 different element names for author, 3 for creator, 22 for contact, 6 for dataset, 5 for identifier, and 13 for name element. These are only representative examples, rather than an exhaustive list.

As mentioned earlier, the descriptive category had the highest counts for element occurrences (Table 3). Most elements in this category tended to be general, e.g., citation, content, data, detail, edition, feature, issue, title, publication, reference, relationship, and source and can be applied across domains. Compared to those in the context category, elements in the descriptive category had shorter names. We further observed that science-oriented elements in geospatial, temporal,

https://doi.org/10.23106/dcmi.952136097

and context categories often had very long element names while general management oriented elements had much shorter names (descriptive, identity, administrative, technical, and semantic).

TABLE 4. Elements that most frequently co-occurred in standards

Element	Occurred in # of standards
Description	10
Title	10
Publisher	8
Citation	8
Country	8
Reference	7
Address	7
Keywords	7
electronicMailAddress	6
Date	6
Abstract	6
Identifier	6
Purpose	6
City	6
Creator	6
postalCode	6
Comment	5
Edition	5
Source	5
Phone	5
Version	5
State or Province	5
Status	5
Name	4
Publication Place	4

4.3 Portability of Elements

The co-occurrence data in Table 4 show that none of the elements occurred in all 16 standards included in this study. The most frequently co-occurred elements are Description and Title, which appeared in 10 standards. The rest of the elements in Table 4 co-occurred in between 4-8 standards. A large number of elements co-occurred in 2-4 standards: 25 elements in 4 standards, 297 elements in 3 standards, and 92 elements in 2 standards. The overall distribution of frequencies of co-occurrences is highly skewed with a very long tail (FIG. 2). This means about one-third (approximately 1500) elements co-occurred in at least two standards while the rest of two-thirds were unique and appeared in only one standard.

A closer examination of the co-occurred elements revealed that, among those co-occurred in two or three standards, most of them fell into related standard clusters. For example, a very large number of elements co-occurred in three standards: the Content Description of Digital Geospatial Metadata (CSDGM) and its two related standards—the Metadata Profile for Shoreline Data and the Biological Data Profile, which are known as the extensions endorsed by CSDGM. Other standards also appeared to be associated through co-occurrences of elements: Darwin Core and ABCD, EML and ABCD, and ClinicalTrial.gov and WHO Trial Registration Data Set. Apparently the close proximity of these disciplinary domains is one of the factors for the element co-occurrences. There were also, however, elements co-occurred in seemingly distant standards. Examples included ABCD and ClinicalTrial.gov, ABCD and Genome Metadata, Biological Data Profile and NISO Metadata for Images in XML. The higher co-occurrences of elements in standards signify their nature of being general in semantics while the lower ones (mainly in the 2 and 3 range) represent the more specific semantic elements in the long tail.

The other measure of metadata element portability is the degree of modularity by examining the substructures that can be independent of a standard and reused by other standards. Since a metadata specification can be implemented on different platforms using varying technical architectures, the element structure in a metadata specification may not be aligned exactly the same way as that in an implementation schema. The data collection for this measure focused only on the standards that have the XML and/or RDF schema available. Among the 16 standards included in this study, we were able to locate XML schemas for only 6 standards. Based on the observation of these schemas, we defined level-1 modularity as having multiple XML schema files for the whole standard and level-2 modularity as having separate schemas for entities such as person/organization, dataset, study, instrument, and subject in addition to level-1 modularity. Almost all standards that made XML schema available achieved level-1 modularity, but level-2 modularity is rare in the standards we observed.

https://doi.org/10.23106/dcmi.952136097

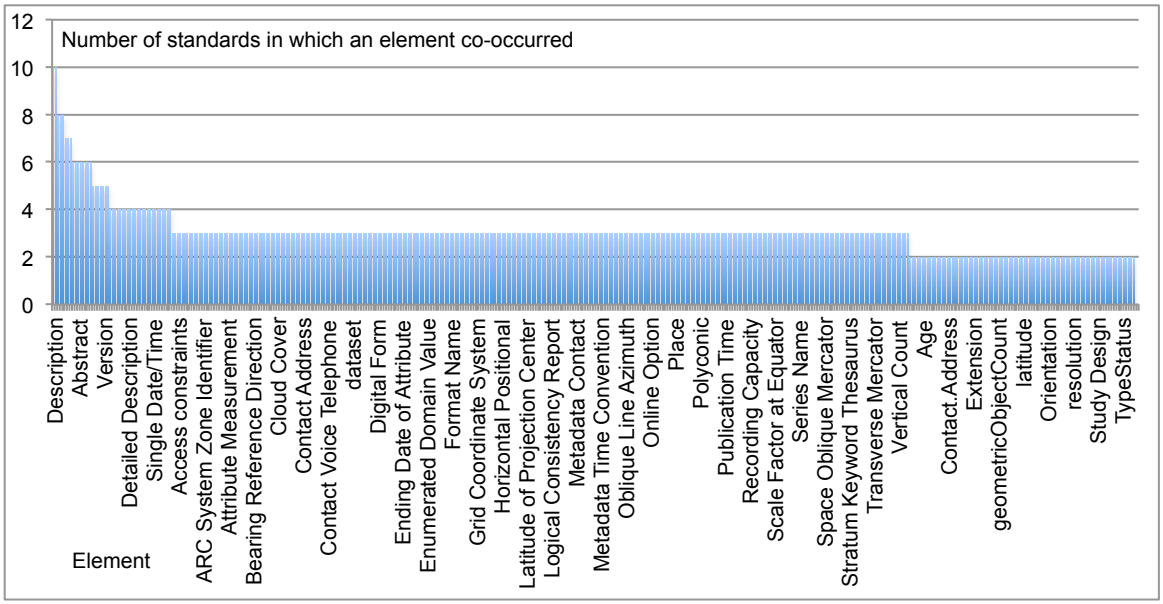

FIG. 2. Frequency distribution of element co-occurrences in science metadata standards

5. Discussion

The survey of metadata standards for scientific data presents some interesting findings that were little known before. Metadata elements in the standards included for this study spread over 9 categories with uneven frequency distributions. Differences in the focus of description caused an extremely high number of elements in the context category. The highest counts of element occurrences fell into the descriptive category and overlapped with DC elements for the most part. This pattern also repeated in the elements co-occurred in different standards. A small number of semantically general elements appeared across the largest numbers of standards while the rest of the element co-occurrences formed a long tail with a wide range of specific semantics. These findings offer valuable insights into the relationship between the portability of metadata elements/schemas and the vision of building a metadata infrastructure for scientific data management, curation, and reuse.

Portability of entity description. Current practice in designing metadata standards is domain-centric and emphasizes one-covers-all that is needed to describe the domain data, regardless whether the same properties of data have had elements established in other standards. This tradition has created a vast number of elements mixed with varying element-naming conventions, which were frequently observed in the process of data processing and coding for this study. The result of such practice is the standards with a very large number of elements and elements with the same semantics but varying greatly in linguistic forms. Not only are such metadata schemes highly complex to design but also difficult to use both by human catalogers and machine processing programs.

As the Semantic Web technologies evolve and application tools and resources are becoming increasingly available, things that were impossible or unrealistic for metadata practice are changing to the positive side. For example, identity standards such as Friend of A Friend (FOAF), ORCID, and DOI provide frameworks for building "identity repositories" or authority name list. They are already serving as the identity metadata schemes for person, organization, dataset, or resource and records for these entities are being created, stored, and shared on open repositories. Identity metadata, therefore, is both technically and theoretically possible to be separated from the metadata standards for scientific data and become an infrastructure service component. So far, at least from the data collected in this study, none of the standards took the

https://doi.org/10.23106/dcmi.952136097

advantage of established identity metadata schemes. The tools for using these standards to create metadata records are mostly standalone and lack the mechanism to utilize the identity resources available. Developing portable entity metadata description will be the first step toward a metadata infrastructure.

Implications for a metadata infrastructure. While more in-depth analysis needs to be done to the metadata elements collected, the findings so far suggest that a standardized, general semantic layer is both necessary and feasible for building a metadata infrastructure. The so-called metadata infrastructure implies that metadata elements, vocabularies, entities, and other metadata artifacts are established as the underlying foundation upon which the tools and applications as well as functions of metadata services are built. Although metadata infrastructure resources have become increasingly available as Semantic Web technologies mature, much of their use is being hindered by the lack of mechanisms for incorporating and customizing them with metadata generation tools. In other words, there is a huge gap between the resources available for metadata creation (vocabularies, authority name lists, geographical names and coordinates, etc.) and the services to utilize these resources as easily as the infrastructure in our daily life. We have taken for granted that a fridge can be hooked to an electrical power outlet and different temperatures for the freeze and fresh food chambers can be set according to our needs. An analogy of a metadata infrastructure would be that a metadata generation tool can interface with the metadata infrastructure services to select the vocabularies, identity instances, geospatial and temporal measurements and units, or any other apparatus needed for generating metadata in a domain. By using infrastructural services, terms, measures, names, and naming conventions are standardized and errors and inconsistencies are minimized. It also simplifies the operation and liberates metadata creation from the highly technical and intellectual processes of metadata element set design and implementation – a fridge user does not need to know its engineering design, so a data librarian does not need to start a metadata project from scratch every time a new research project generates new data.

The portability of metadata elements is closely tied with a metadata infrastructure, which implies more than modules in metadata standards. Using identity metadata as an example, a data librarian may utilize the metadata infrastructure services to customize a metadata scheme for a research team with built-in identity metadata for the team members to reduce manual data entry and human errors. Similarly, semantic, geospatial, and temporal metadata should also be able to leverage the established infrastructural resources to customize metadata generation tools or programs. All these possibilities are not only possible, but should also be the direction for metadata applications in the science data domain. The one-covers-all style in current metadata standards, however, will make it very difficult, if not impossible, for a significant breakthrough in keeping up with the exponential growth of scientific data.

6. Conclusion and Future Research

This paper reported findings from a survey to metadata standards in the scientific data domain and argued for a metadata infrastructure. Being portable is the essential condition or prerequisite for metadata schemes to be "infrastructurized" – a word we coined to denote the state of being built into or as part of the infrastructure. It is also one of the three principles (least effort, portability, and infrastructure) for fulfilling the metadata functional requirements (Qin, Ball, & Greenberg, 2012).

As much as we are excited about the idea of metadata infrastructure and its potential for radically changing the style and approach that metadata operates, there are still many research questions to be addressed. What should a metadata infrastructure constitute? How can the gaps be filled or narrowed between the infrastructure resources and metadata applications? Is it possible or is there a need to streamline the metadata scheme design practice toward a metadata infrastructure? The list can go on. The bottomline here is that technologies have advanced to a

https://doi.org/10.23106/dcmi.952136097

point that made the old metadata paradigm outdated. We are facing both tremendous challenges and opportunities to make the best out of these challenges.

The data collected in our survey offered some insights about the current status of metadata scheme design in the scientific data. The next step will be to continue the pattern analysis for the element data and generalize the semantic and linguistic pattern elements for elements common across standards. We are also interested in experimenting with these generalized elements for "infrasturcturization" by using Semantic Web technologies.

References

caGrid. (2011). Metadata 1.4 Documentation. Retrieved March 31, 2013 from http://www.cagrid.org/display/metadata14/Documentation.

Dunlap, Rocky, Leo Mark, Spencer Rugaber, V. Balaji, Julien Chastang, Luca Cinquini, Cecelia DeLuca, Don Middleton, and Sylvia Murphy. (2008). Earth system curator: Metadata infrastructure for climate modeling. Earth Science Informatics, 1: 131-149.

Godby, Jean, Devon Smith, and Eric Childress. (2003). Two paths to interoperable metadata. Proceedings of the International Conference for Dublin Core and Metadata Applications 2003. Retrieved March 31, 2013 from https://www.oclc.org/content/dam/research/publications/library/2003/godby-dc2003.pdf.

Jones, Matthew B., Chad Berkley, Jivka Bojilova,,and Mark Schildhauer. (2001). Managing scientific metadata. *IEEE* Internet Computing,5 (5).

Michener, W. K. (2006). Meta-information concepts for ecological data management. Ecological Informatics, 1 (1), 3-7.

Millerand, F. & Baker, K.S. (2010). Who are the users? Who are the developers? Webs of users and developers in the development process of a technical standard. Information Systems Journal, 20: 137-161.

Qin, Jian, Alex Ball, and Jane Greenberg. (2012). Functional and architectural requirements for metadata: Supporting discovery and management of scientific data. Proceedings of the International Conference on Dublin Core and Metadata Applications 2012. Retrieved March 31, 2013 from http://dcpapers.dublincore.org/pubs/article/view/3660/1883

Riall, Rebecca L., Fausto Marincioni, and Frances L. Lightsom. (2004). Content metadata standards for marine science: A case study: Chapter: Evolution. USGS Open-File Report 2004-1002. Retrieved, March 31, 2013, from http://pubs.usgs.gov/of/2004/1002/html/evol.html.

Star, S.L. & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information space. Information Systems Research, 7(1): 111-134.

Tennant, Roy. (2004). A bibliographic metadata infrastructure for the twenty-first century. Library Hi Tech, 22(2): 175-181.

U.S. RDA Testing Coordinating Committee. (2013). Final U.D. RDA implementation update (January 4, 2013). Retrieved, March 31, 2013, from http://www.loc.gov/aba/rda/pdf/RDA_updates_04jan13.pdf.

Willis, Craig, Jane Greenberg, and Hollie White. (2012). Analysis and synthesis of metadata goals for scientific data. Journal of the American Society for Information Science and Technology, 63(8): 1505-1520.

Wittenburg, Peter. (2009). Metadata infrastructure for Language Resources and Technology. 2009-02-04 – Version 5. Retrieved March 31, 2013 from http://www.clarin.eu/sites/default/files/wg2-4-metadata-doc-v5.pdf

https://doi.org/10.23106/dcmi.952136097